

# Full likelihood inference for max-stable data

Clément Dombry<sup>1</sup>, Marc G. Genton<sup>2</sup>, Raphaël Huser<sup>2</sup>, Mathieu Ribatet<sup>3</sup>

March 28, 2017

## Abstract

We show how to perform full likelihood inference for max-stable multivariate distributions or processes based on a stochastic Expectation-Maximisation algorithm. In contrast to current approaches, such as pairwise likelihoods or the Stephenson–Tawn likelihood, our method combines statistical and computational efficiency in high-dimensions, and it is not subject to bias entailed by lack of convergence of the underlying partition. The good performance of this methodology is demonstrated by simulation based on the logistic model, and it is shown to provide dramatic computational time improvements with respect to a direct computation of the likelihood. Strategies to further reduce the computational burden are also discussed.

**Keywords:** Full likelihood; Max-stable distribution; Stephenson–Tawn likelihood; Stochastic EM algorithm.

---

<sup>1</sup>Department of Mathematics, University of Franche-Comté, 25030 Besançon cedex, France. E-mail: clement.dombry@univ-fcomte.fr

<sup>2</sup>CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: marc.genton@kaust.edu.sa and raphael.huser@kaust.edu.sa

<sup>3</sup>Department of Mathematics, University of Montpellier, 34095 Montpellier cedex 5, France. E-mail: mathieu.ribatet@umontpellier.fr

# 1 Introduction

Provided one believes in asymptotics and that some mild regularity conditions are met, max-stable distributions and processes can be useful for studying high-dimensional extreme events recorded in space and/or time (Padoan *et al.*, 2010; Davis *et al.*, 2013; de Carvalho and Davison, 2014; Huser and Davison, 2014; Embrechts *et al.*, 2015; Huser and Genton, 2016). This broad but constrained class of models may, at least theoretically, be used to extrapolate into the joint tail, hence providing a justified framework for risk assessment of extreme events. The probabilistic justification is that the max-stable property arises in limiting models for suitably rescaled maxima of independent and identically distributed processes. For recent reviews, see Davison *et al.* (2012) and Davison and Huser (2015).

Because extremes are rare by definition, it is crucial for reliable estimation and prediction to extract as much information from the data as possible. Thus, efficient estimators play a particularly important role in statistics of extremes, and the maximum likelihood estimator (MLE) is a natural choice thanks to its appealing large-sample properties. However, the likelihood function is excessively difficult to compute for high-dimensional max-stable data. As detailed in §3, likelihood evaluations require the computation of a sum indexed by all partitions  $\pi$  in a given set  $\mathcal{P}_D$ , the cardinality of which grows more than exponentially with the dimension,  $D$ . In a thorough simulation study, Castruccio *et al.* (2016) stated that current technologies are limiting full likelihood inference to dimension 12 or 13, and they concluded that without meaningful methodological advances, a direct full likelihood approach will not be feasible.

To circumvent this computational bottleneck, several strategies have been advocated. Padoan *et al.* (2010) proposed a pairwise likelihood approach, combining the bivariate densities of carefully chosen pairs of observations. Although this method is computationally attractive and inherits many good properties from the MLE, it also entails a loss in efficiency, which becomes more apparent in high dimensions (Huser *et al.*, 2016). More efficient triplewise and higher-order composite likelihoods were investigated by Genton *et al.* (2011),

Huser and Davison (2013) and Castruccio *et al.* (2016). However, they are still not fully efficient, and it is not clear how to optimally select the composite likelihood components. Furthermore, because composite likelihoods are generally not valid likelihoods (Varin *et al.*, 2011), the classical likelihood theory cannot be blindly applied for uncertainty assessment, testing, model validation and selection, and so forth.

Alternatively, Stephenson and Tawn (2005) suggested augmenting the  $n$ -block maxima data  $z^n = (z_1^n, \dots, z_D^n)^T$  with their occurrence times, which determine an observed partition  $\pi^n$  of the set  $\{1, \dots, D\}$ . Loosely speaking, assuming that  $\pi^n$  is well approximated by its limit partition  $\pi$  and that  $z^n$  is approximately max-stable, this extra information yields a drastically simplified joint likelihood for  $z^n$  and  $\pi^n$ . However, Wadsworth (2015) and Huser *et al.* (2016) noted that lack of convergence of  $\pi^n$  to  $\pi$  may lead to severe bias, especially in low-dependence scenarios. By fixing  $\pi$  to the observed partition  $\pi^n$ , a strong constraint is imposed, creating model misspecification, to which likelihood methods are very sensitive. A related approach is to take advantage of the limiting Poisson point process representation of extremes, yielding efficient inference methods based on a variety of threshold-based likelihoods (see Huser *et al.*, 2016). In particular, the censored Poisson likelihood and Stephenson–Tawn likelihood coincide when the marginal thresholds are taken to be the observed maxima  $z^n$  (Wadsworth and Tawn, 2014). Thus, their efficiency and robustness properties are similar (Huser *et al.*, 2016).

In this paper, to avoid bias entailed by fixing the partition, we suggest returning to the original likelihood formulation, which integrates out the partition rather than conditioning on it. By interpreting the partition as a missing observation, we show how to design a stochastic Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977; Nielsen, 2000) for efficient inference. The quality of the stochastic approximation to the full likelihood can be controlled and set to any arbitrary precision at a computational cost. We show that much higher-dimensional max-stable models may be fitted in reasonable time. Importantly, our method is based solely on max-stable data and does not require additional information

about the partition or the original processes, unlike the Stephenson–Tawn and threshold-based likelihoods. Our approach is based on the algorithm of [Dombry \*et al.\* \(2013\)](#) for conditional simulation of the partition given the data, and it is closely related to the recent contribution of [Thibaud \*et al.\* \(2016\)](#), who in a Bayesian setting developed a Markov chain Monte Carlo algorithm for a specific class of max-stable processes by treating the partition as a latent variable to be resampled at each iteration.

## 2 Max-stable processes and distributions

### 2.1 Definition, construction, and models

Consider a sequence of independent and identically distributed processes  $Y_1(s), Y_2(s), \dots$ , indexed by spatial location  $s \in \mathcal{S} \subset \mathbb{R}^d$ , and assume that there exist sequences of functions  $a_n(s) > 0$  and  $b_n(s)$ , such that the renormalised  $n$ -block maximum process

$$Z^n(s) = a_n(s)^{-1} [\max\{Y_1(s), \dots, Y_n(s)\} - b_n(s)] \quad (1)$$

converges in distribution to a process  $Z(s)$  with non-degenerate margins, i.e.,  $Y(s)$  is in the max-domain of attraction of  $Z(s)$ . Then, the limit  $Z(s)$  is max-stable (see, e.g., [de Haan and Ferreira, 2006](#), Chap. 9) in the sense that when the processes  $Y_1(s), \dots, Y_n(s)$  in (1) are substituted by  $n$  independent copies of  $Z(s)$ , then the  $n$ -block maximum process  $Z^n(s)$  is *equal* in distribution to  $Z(s)$ , for any integer  $n$ . By definition of a max-stable process, all finite-dimensional distributions are max-stable. In particular, univariate margins are generalised extreme-value distributed.

Consider now points of a unit rate Poisson point process,  $P_1, P_2, \dots$ , and independent copies,  $W_1(s), W_2(s), \dots$ , of a positive stochastic process  $W(s)$  with unit mean. Then, the process

$$Z(s) = \sup_{j \geq 1} W_j(s)/P_j, \quad s \in \mathcal{S}, \quad (2)$$

is max-stable with unit Fréchet margins, i.e.,  $\Pr\{Z(s) \leq z\} = \exp(-1/z)$ ,  $z > 0$  ([de Haan, 1984](#); [Schlather, 2002](#)). Representation (2) provides a way to build a wide variety of max-

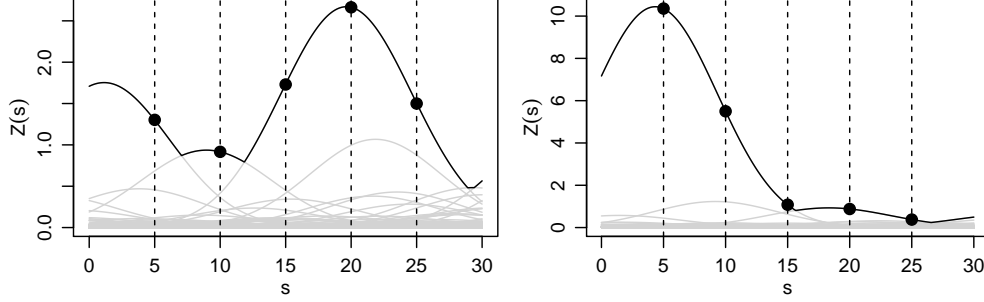


Figure 1: Two realisations (black) from the same [Smith \(1990\)](#) max-stable process on the line defined by setting  $W(s) = \phi(s - U; \sigma^2)$ ,  $s \in \mathcal{S} = \mathbb{R}$ , in (2), with the latent profiles  $W_j(s)/P_j$  (grey). Here,  $\sigma^2 = 5$ . When observed at locations 5, 10, 15, 20, 25, the partitions are  $\pi = \{\{1\}, \{2\}, \{3, 4, 5\}\}$  (left) and  $\pi = \{\{1, 2, 3\}, \{4, 5\}\}$  (right).

stable processes ([Smith, 1990](#); [Schlather, 2002](#); [Kabluchko et al., 2009](#); [Opitz, 2013](#)) and multivariate distributions if  $\mathcal{S}$  is finite, and to simulate from them ([Schlather, 2002](#); [Dombry et al., 2016](#)). Furthermore, from (2), one can deduce that the joint distribution at the sites  $s_1, \dots, s_D \in \mathcal{S}$  is

$$\Pr\{Z(s_1) \leq z_1, \dots, Z(s_D) \leq z_D\} = \exp\{-V(z_1, \dots, z_D)\}, \quad (3)$$

where the function  $V(z_1, \dots, z_D) = \mathbb{E}[\max\{W(s_1)/z_1, \dots, W(s_D)/z_D\}]$  is the *exponent measure* and satisfies certain constraints (see, e.g., [Davison and Huser, 2015](#)). As an illustration, Figure 1 shows two independent simulations from the same [Smith \(1990\)](#) model defined by taking  $W(s) = \phi(s - U; \sigma^2)$ ,  $s \in \mathcal{S} = \mathbb{R}$ , in (2), where  $\phi(\cdot; \sigma^2)$  is the normal density with zero mean and variance  $\sigma^2$ , and  $U$  is a point from a unit rate Poisson point process on the line.

## 2.2 Underlying partition and extremal functions

At each location  $s \in \mathcal{S}$ , the pointwise supremum,  $Z(s)$ , in (2) is realised by a single profile  $W_j(s)/P_j$  almost surely. Such profiles are called extremal functions in [Dombry et al. \(2013\)](#). As illustrated in Figure 1, the extremal functions are only partially observed and define a latent random partition  $\pi = \{\tau_1, \dots, \tau_k\}$  of the set  $\{1, \dots, D\}$  (also called hitting scenario in [Dombry et al., 2013](#)), identifying clusters of variables stemming from the same event. For

example, on the left-hand panel of Figure 1, the partition  $\pi = \{\{1\}, \{2\}, \{3, 4, 5\}\}$  indicates that the max-stable process at these five locations essentially came from three separate independent events; in particular the maxima at locations 3, 4, and 5 were generated from the same profile.

Similarly, an observed partition  $\pi^n$  of  $\{1, \dots, D\}$  may be defined for  $Z^n(s)$  in (1) using the original processes  $Y_1(s), \dots, Y_n(s)$ . The knowledge of  $\pi^n$  tells us if extreme events at different locations occurred together or not, and it therefore contains some information about the strength of spatial extremal dependence. If the data  $Y_1(s), \dots, Y_n(s)$  are in the max-domain of attraction of  $Z(s)$  defined by (2), then the partition  $\pi^n$  converges in distribution to  $\pi$  on the space of all partitions  $\mathcal{P}_D$  of  $\{1, \dots, D\}$  (Stephenson and Tawn, 2005; Wadsworth, 2015).

We now describe likelihood inference for max-stable vectors: by conditioning on the partition  $\pi$  (so-called *Stephenson–Tawn likelihood*), or by integrating it out (so-called *full likelihood*).

### 3 Likelihood inference

#### 3.1 Full and Stephenson–Tawn likelihoods

By differentiating the distribution (3) with respect to the variables  $z_1, \dots, z_D$ , one can deduce that the corresponding density, or full likelihood for one replicate, may be expressed as

$$g_{\text{Full}}(z_1, \dots, z_D) = \exp\{-V(z_1, \dots, z_D)\} \sum_{\pi=\{\tau_1, \dots, \tau_k\} \in \mathcal{P}_D} \prod_{i=1}^k \{-V_{\tau_i}(z_1, \dots, z_D)\}, \quad (4)$$

where  $V_{\tau_i}$  denotes the partial derivative of the function  $V$  with respect to the variables indexed by the set  $\tau_i \subset \{1, \dots, D\}$  (Huser *et al.*, 2016; Castruccio *et al.*, 2016). The sum in (4) is taken over all elements of  $\mathcal{P}_D$ , the size of which equals the Bell number, leading to an explosion of terms for large  $D$ . Each term in (4) corresponds in fact to a different configuration of the profiles  $W_j(s)/P_j$  in (2) at the sites  $s_1, \dots, s_D \in \mathcal{S}$ . For this reason, Castruccio *et al.* (2016) concluded that the computation of (4) is limited to dimension 12 or 13 with modern computational resources.

As detailed in the online Supplementary Material using a point process argument based on extremal functions, and originally shown by [Stephenson and Tawn \(2005\)](#), if all profiles  $W_j(s)/P_j$  or at least the partition  $\pi$  were observed, the *joint* density for the max-stable data  $z = (z_1, \dots, z_D)^T$  and the partition  $\pi = \{\tau_1, \dots, \tau_k\}$  would simply be

$$g_{\text{ST}}(z_1, \dots, z_D, \pi) = \exp\{-V(z_1, \dots, z_D)\} \prod_{i=1}^k \{-V_{\tau_i}(z_1, \dots, z_D)\}, \quad (5)$$

hence reducing the problematic sum to a single term, making likelihood inference possible and simultaneously improving statistical efficiency. Because the asymptotic partition  $\pi$  is not observed, [Stephenson and Tawn \(2005\)](#) suggested replacing it by the observed partition  $\pi^n$  of occurrence times of maxima, which converges to  $\pi$  provided the asymptotic model is well specified. This ingenious idea was implemented, for example, by [Davison and Gholamrezaee \(2012\)](#) in a study of extreme temperatures in Switzerland. However, as underlined by [Wadsworth \(2015\)](#) and [Huser \*et al.\* \(2016\)](#), lack of convergence of  $\pi^n$  to  $\pi$  may result in severe estimation bias, which is especially strong in low-dependence cases, a scenario frequently encountered in practice. To circumvent this problem, [Wadsworth \(2015\)](#) proposed a bias-corrected likelihood; alternatively, we show in the next subsection how to design a stochastic EM algorithm to maximise (4), while taking advantage of the computationally appealing nature of (5).

### 3.2 Stochastic Expectation-Maximisation algorithm

It is instructive to rewrite the full likelihood (4) using (5) as  $g_{\text{Full}}(z_1, \dots, z_D) = \sum_{\pi \in \mathcal{P}_D} g_{\text{ST}}(z_1, \dots, z_D, \pi)$  because it clearly shows that the full likelihood simply integrates out the latent random partition  $\pi$  needed for the Stephenson–Tawn likelihood. Interpreting  $\pi$  as a missing observation and the Stephenson–Tawn likelihood as the completed likelihood, an EM algorithm ([Dempster \*et al.\*, 1977](#)) may be easily formulated. Assume that the exponent measure  $V(z_1, \dots, z_D \mid \theta)$  is parametrised by a vector  $\theta \in \Theta \subset \mathbb{R}^p$ . Starting from an initial guess  $\theta_0 \in \Theta$ , the EM algorithm consists of iterating the following E- and M-steps for  $r = 1, \dots, R$ :

- E-step: compute the functional

$$Q(\theta, \theta_{r-1}) = \mathbb{E}_{\pi|z, \theta_{r-1}} [\log \{g_{\text{ST}}(z, \pi | \theta)\}] = \sum_{\pi \in \mathcal{P}_D} g(\pi | z, \theta_{r-1}) \log \{g_{\text{ST}}(z, \pi | \theta)\}, \quad (6)$$

where the expectation is computed with respect to the discrete conditional distribution of  $\pi$  given the data  $z = (z_1, \dots, z_D)^T$  and the current value of the parameter  $\theta_{r-1}$ , i.e.,

$$g(\pi | z, \theta_{r-1}) = g_{\text{ST}}(z, \pi | \theta_{r-1}) / g_{\text{Full}}(z | \theta_{r-1}). \quad (7)$$

- M-step: update the parameter as  $\theta_r = \arg \max_{\theta \in \Theta} Q(\theta, \theta_{r-1})$ .

[Dempster \*et al.\* \(1977\)](#) showed that the EM algorithm has good properties; in particular, the value of the log-likelihood increases at each iteration, which ensures convergence of  $\theta_r$  to a local maximum, as  $r \rightarrow \infty$ . In our case, however, the expectation in (6) is tricky to compute: it contains again the sum over the set  $\mathcal{P}_D$ , and (7) relies on the full density  $g_{\text{Full}}(z | \theta_{r-1})$ , which we try to avoid. To circumvent this issue, one solution is to stochastically approximate (6) by

$$\hat{Q}(\theta, \theta_{r-1}) = \frac{1}{N} \sum_{i=1}^N \log \{g_{\text{ST}}(z, \pi_i | \theta)\}, \quad \pi_1, \dots, \pi_N \sim g(\pi | z, \theta_{r-1}), \quad (8)$$

where the conditional partitions  $\pi_1, \dots, \pi_N$  are independent at best, or form an ergodic sequence at least. As  $g(\pi | z, \theta_{r-1}) \propto g_{\text{ST}}(z, \pi | \theta_{r-1})$ , see (7), it is possible to devise a Gibbs sampler to efficiently simulate after some burn-in an ergodic sequence of draws from  $g(\pi | z, \theta_{r-1})$  *without* explicitly computing the constant factor  $g_{\text{Full}}(z | \theta_{r-1})$  in the denominator of (7). Thanks to ergodicity of the resulting Markov chain, the precision of the approximation (8) may be set arbitrarily high by letting  $N \rightarrow \infty$ . For more details about the implementation of the Gibbs sampler see [Dombry \*et al.\* \(2013\)](#). Notice that although the number of iterations of the Gibbs sampler,  $N$ , will typically be much smaller than the cardinality of  $\mathcal{P}_D$ , the approximation (8) to (6) will likely be reasonably good for moderate values of  $N$  because only a few partitions  $\pi \in \mathcal{P}_D$  may be plausible or compatible with the data  $z = (z_1, \dots, z_D)^T$ . The asymptotic properties of the stochastic EM estimator,  $\hat{\theta}_{\text{SEM}}$ , were studied in details by [Nielsen \(2000\)](#) and compared with the classical MLE,  $\hat{\theta}$ .



Consistency and asymptotic normality of  $\widehat{\theta}_{\text{SEM}}$  can be established under mild regularity conditions, and the asymptotic performance is very similar to that of  $\widehat{\theta}$ . Furthermore, the inherent variability of the stochastic EM algorithm is also a blessing: unlike the deterministic EM algorithm, it is less likely to get stuck at a local maximum of the full likelihood.

## 4 Simulation study

To assess the performance of the stochastic EM algorithm, we simulate data from the multivariate logistic max-stable distribution defined by its exponent measure  $V(z_1, \dots, z_D \mid \theta) = (\sum_{j=1}^D z_j^{-1/\theta})^\theta$ ,  $\theta \in \Theta = (0, 1]$ . Here, the parameter  $\theta$  controls the dependence strength, with  $\theta \rightarrow 0$  and  $\theta = 1$  corresponding to perfect dependence and independence, respectively. This model was chosen for two main reasons: first, it is the simplest, non-trivial, and most widely-used max-stable distribution, often used as a benchmark, and second, the full likelihood (4) can be efficiently computed in this case using a recursive algorithm (Shi, 1995), thus allowing us to compare  $\widehat{\theta}_{\text{SEM}}$  and  $\widehat{\theta}$  in high dimensions.

We first investigated the performance of  $\widehat{\theta}_{\text{SEM}}$  under different scenarios. We considered dimensions  $D = 2, 5, 10, 20$  with 20 independent temporal replicates, and  $\theta = 0.1, \dots, 0.9$  (strong to weak dependence). Setting the initial value to  $\theta_0 = 0.6$ , we chose  $R = 30$  iterations for the EM algorithm, averaging the last 5 iterations, and took  $100 \times D$  iterations for the underlying Gibbs sampler. Thinning by a factor  $D$  was applied to keep  $N = 100$  roughly independent partitions  $\pi_i$  to compute (8). We repeated the experiment 1000 times to estimate the bias, B, standard deviation, SD, root mean squared error,  $\text{RMSE} = (\text{B}^2 + \text{SD}^2)^{1/2}$ , and absolute relative error with respect to the MLE,  $\text{RE} = \text{E}|(\widehat{\theta}_{\text{SEM}} - \widehat{\theta})/\widehat{\theta}|$ . Figure 2 reports the results. As expected, the bias is negligible with respect to the standard deviation, and the latter decreases with increasing dimension  $D$  but increases as the data approach independence ( $\theta \rightarrow 1$ ). The root mean squared error is almost only determined by the standard deviation. The relative error (not shown) is always very small (uniformly less than 0.5%), and simulations suggest that it decreases with  $D$  and  $N$ .

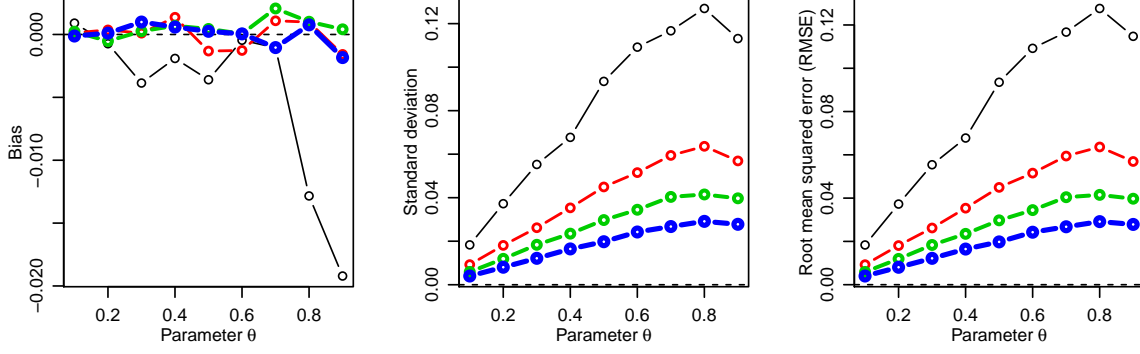


Figure 2: Performance of the stochastic EM estimator  $\hat{\theta}_{\text{SEM}}$ : bias (left), standard deviation (middle) and root mean squared error (right), for data distributed according to the logistic model with  $\theta = 0.1, \dots, 0.9$  in dimension  $D = 2$  (thinnest black), 5 (thin red), 10 (thick green) and 20 (thickest blue), based on 20 independent temporal replicates. The number of iterations for the EM algorithm was set to  $R = 30$ , averaging the last 5 iterations, and the number of iterations of the underlying Gibbs sampler was set to  $100 \times D$ . Thinning by a factor  $D$  was applied to keep  $N = 100$  roughly independent partitions  $\pi_i$  to compute (8). The initial value was set to  $\theta_0 = 0.6$ .

We now turn our attention to the computational efficiency of the stochastic EM algorithm. Considering dimensions up to  $D = 100$  under the exact same setting as before, the leftmost panel of Figure 3 shows that it takes approximately half a day to compute  $\hat{\theta}_{\text{SEM}}$  when  $D = 100$  and  $\theta = 0.9$ . Recall that, according to [Castruccio \*et al.\* \(2016\)](#), a direct evaluation of the likelihood (4) is not possible in dimensions greater than  $D = 12$  or 13, thus this result is a great improvement over the current existing methods. However, because it is still fairly intensive to compute, it makes sense to seek strategies to reduce the computational burden. One possibility is to tune the number of iterations of the stochastic EM algorithm. To investigate its speed of convergence, the two middle panels of Figure 3 show the sample path  $r \mapsto \theta_r$  as a function of the EM iteration  $r = 1, \dots, 50$ , centred by their average over iterations 30–50 for 100 independent runs in dimension  $D = 10$ . The true values were set to  $\theta = 0.3$  (second panel) and 0.9 (third panel), and the initial value was set to  $\theta_0 = 0.6$ . Note that the convergence is quite fast when  $\theta = 0.3$ , requiring about 5 iterations, but when  $\theta = 0.9$ , it takes about 15–30 iterations. Further simulations (not shown) suggest that about 5 iterations are enough for the algorithm to converge in all cases when the initial value is

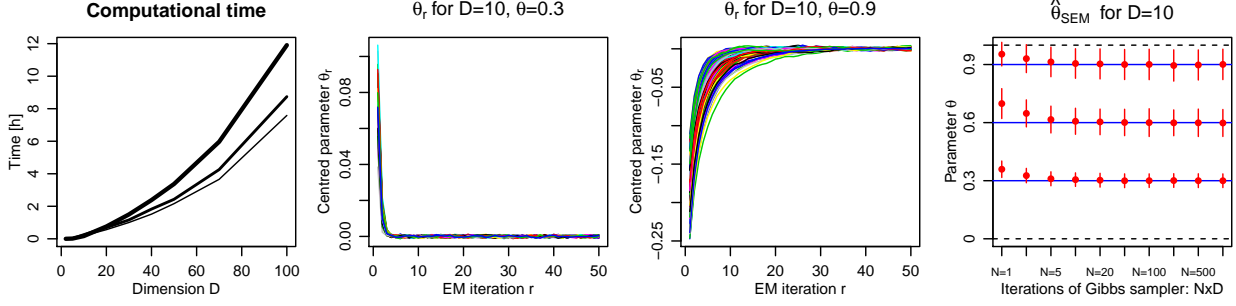


Figure 3: Left: Computational time of  $\hat{\theta}_{SEM}$  as function of dimension  $D$ , for  $\alpha = 0.3$  (thin),  $0.6$  (medium),  $0.9$  (thick). We used 20 temporal replicates, 30 EM iterations, and  $100 \times D$  iterations for the underlying Gibbs sampler. Middle panels: parameter values  $\theta_r$  as function of EM iteration  $r = 1, \dots, 50$ , centred by the average over iterations 30–50, for 100 independent runs in dimension  $D = 10$ . True values were set to  $\theta = 0.3$  (second panel) and  $0.9$  (third panel), and the initial value was set to  $\theta_0 = 0.6$ . Right: Mean of estimated parameters  $\hat{\theta}_{SEM}$  (red dots) with 95% confidence intervals for  $\theta = 0.3, 0.6, 0.9$ , dimension 10, 30 EM iterations, and  $N \times D$  Gibbs sampler iterations with  $N = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000$  (x-axis).

adequately chosen, for example using a pairwise likelihood estimator, and that these results are similar for other dimensions. For other models with a higher-dimensional parameter space (say  $p = 2-4$ ), we expect the required number of EM iterations to be slightly larger but to stay fairly small provided good starting values may be chosen. Another possibility to reduce the computational time is to play with the number of iterations of the underlying Gibbs sampler, which is the main computational bottleneck of this approach. The rightmost panel of Figure 3 displays estimated parameters in dimension  $D = 10$  with associated 95% confidence intervals for  $\theta = 0.3, 0.5, 0.9$ , using 30 EM iterations and  $N \times D$  iterations for the underlying Gibbs sampler with  $N = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000$ . Thinning by a factor  $D$  was applied to keep  $N$  effective iterations used in (8). Surprisingly, the distribution of  $\hat{\theta}_{SEM}$  is almost stable for  $N \geq 5-10$ , suggesting that the number of Gibbs iterations does not need to be very large for accurate estimation. Overall, as the complexity is roughly linear in  $R$  (EM iterations) and  $N$  (Gibbs iterations), computing  $\hat{\theta}_{SEM}$  when  $D = 100$  and  $\theta = 0.9$  may easily be reduced from 12 hr to less than an hour, if  $R$  and  $N$  are properly chosen, without any loss of accuracy.

## 5 Discussion

To resolve the problem of inference for max-stable distributions and processes, we have proposed a stochastic EM algorithm, which does not fix the underlying partition, but instead, treats it as a missing observation, and integrates it out. The beauty of this approach is that it combines statistical and computational efficiency in high dimensions, and it does not suffer from misspecification entailed by lack of convergence of the partition. As a proof of concept, we have validated the methodology by simulation based on the logistic model, and we have shown that it is fairly easy to make inference up to dimension  $D = 100$ , which could likely be pushed to  $D = 300$ – $500$  for the logistic model. However, our method is not limited to the logistic model and could potentially be applied to any max-stable model for which the exponent measure  $V$  and its partial derivatives  $V_{\tau_i}$  are available. For more realistic models, such as the Brown–Resnick model ([Kabluchko \*et al.\*, 2009](#)), the difficulty resides in the computation of high-dimensional Gaussian multivariate distributions needed for  $V$  and  $V_{\tau_i}$ . Unbiased Monte Carlo estimates of these quantities can be obtained, and [Thibaud \*et al.\* \(2016\)](#) and [de Fondeville and Davison \(2016\)](#) suggest using crude approximations to reduce the computational time while maintaining accuracy. For Brown–Resnick-like models, we therefore expect to be able to handle up to dimensions  $D = 50$ – $100$  in a reasonable amount of time. The main computational bottleneck of our approach is that we need to generate a Gibbs sampler for each independent temporal replicate of the process. Fortunately, as this setting is embarrassingly parallel, one may thus easily take advantage of available distributed computing resources. Finally, there is a large volume of literature on the stochastic EM algorithm, and it might be possible to devise automatic stopping criteria and adaptive schemes for the Gibbs sampler to further speed up the algorithm ([Booth and Hobert, 1999](#)).

# A Supplementary material

## A.1 Likelihood derivation via Poisson point process intensity

In their original paper, [Stephenson and Tawn \(2005\)](#) derived the likelihood  $g_{\text{Full}}$  and  $g_{\text{ST}}$  by differentiating the cumulative distribution function

$$\Pr\{Z(s_1) \leq z_1, \dots, Z(s_D) \leq z_D\} = \exp\{-V(z_1, \dots, z_D)\}.$$

Here, we propose a different approach based on the analysis of the Poisson point process representation of the max-stable process

$$Z(s) = \sup_{j \geq 1} W_j(s)/P_j, \quad s \in \mathcal{S}. \quad (9)$$

Introducing the functions  $\varphi_j = W_j/P_j$ ,  $j = 1, 2, \dots$ , the point process  $\Phi = \{\varphi_j, j \geq 1\}$  is a Poisson point process on the space of nonnegative functions defined on  $\mathcal{S}$ . The max-stable process  $Z$  appears as the pointwise maximum of the functions in  $\Phi$ . [Dombry and Éyi-Minko \(2013\)](#) showed that for all locations  $s \in \mathcal{S}$ , there almost surely exists a unique function in  $\Phi$  that reaches the maximum  $Z(s)$  at  $s$ . This function is called the *extremal function* at  $s$  and denoted by  $\varphi_s^+$ . Clearly,  $Z(s) = \varphi_s^+(s)$ .

Given  $D$  locations  $s_1, \dots, s_D \in \mathcal{S}$ , there can be repetitions within the extremal functions  $\varphi_{s_1}^+, \dots, \varphi_{s_D}^+$ , meaning that the maximum at different locations  $s_{j_1}, s_{j_2}$ , can arise from the same extremal event. The notion of *hitting scenario* accounts for such possible repetitions. It is defined as the random partition  $\pi = \{\tau_1, \dots, \tau_k\}$  of  $\{1, \dots, D\}$  such that the two indices  $j_1$  and  $j_2$  are in the same block if and only if the extremal functions at  $s_{j_1}$  and  $s_{j_2}$  are equal. Here  $k$  denotes the number of blocks of the partition  $\pi$  and is equal to the number of different functions in  $\Phi$  reaching the maximum  $Z(s)$  for some point  $s \in \{s_1, \dots, s_D\}$ . Within the block  $\tau_i$ , all the points  $s_j$ ,  $j \in \tau_i$ , share the same extremal function that will hence be denoted by  $\varphi_{\tau_i}^+$ .

The joint distribution of the hitting scenario  $\pi = \{\tau_1, \dots, \tau_k\}$  and extremal functions  $\{\varphi_{\pi_1}^+, \dots, \varphi_{\pi_k}^+\}$  was derived by [Dombry and Éyi-Minko \(2013\)](#). The max-stable observations

$Z(s_1), \dots, Z(s_D)$  relate to the hitting scenario and extremal functions via the simple equation  $Z(s_j) = \varphi_{\tau_i}^+(s_j)$  for  $j \in \tau_i$ . In this way, we can deduce the joint distribution of the partition  $\pi = \{\tau_1, \dots, \tau_k\}$  and max-stable observations  $Z(s_1), \dots, Z(s_D)$ , i.e., the Stephenson–Tawn likelihood  $g_{\text{ST}}$ . Marginalising out the random partition, we deduce the full likelihood  $g_{\text{Full}}$ .

Suppose that the random vectors  $\{W_j(s_1), \dots, W_j(s_D)\}^T$ ,  $j \geq 1$ , stemming from (9), have a density  $f_W$  with respect to the Lebesgue measure on  $(0, +\infty)^D$ . Then, the Poisson point process  $\{\{\varphi_j(s_1), \dots, \varphi_j(s_D)\}^T, j \geq 1\}$  on  $(0, +\infty)^D$  has intensity

$$\lambda(z_1, \dots, z_D) = \int_0^\infty f_W(z_1/r, \dots, z_D/r) r^{-2-D} dr. \quad (10)$$

For clarity, we introduce some vectorial notation: let  $\mathbf{s} = (s_1, \dots, s_D)^T$ ,  $\mathbf{z} = (z_1, \dots, z_D)^T$ ,  $Z(\mathbf{s}) = \{Z(s_1), \dots, Z(s_D)\}^T$ . If  $\tau_i$  is a subset of  $\{1, \dots, D\}$ ,  $\tau_i^c$  denotes the complementary subset, while  $\mathbf{z}_{\tau_i}$  and  $\mathbf{z}_{\tau_i^c}$  denote the subvectors of  $\mathbf{z}$  obtained by keeping only the components from  $\tau_i$  and  $\tau_i^c$ , respectively. Proposition 3 in [Dombry and Éyi-Minko \(2013\)](#) yields the following results:

- From the Poisson point process property, one can deduce the joint law of the hitting scenario and extremal functions:

$$\Pr\{\pi = \{\tau_1, \dots, \tau_k\}, \varphi_{\tau_1}^+(\mathbf{s}) = d\mathbf{z}_1, \dots, \varphi_{\tau_k}^+(\mathbf{s}) = d\mathbf{z}_k\} = \exp\{-V(\max_{i=1}^k \mathbf{z}_i)\} \prod_{i=1}^k \lambda(\mathbf{z}_i) d\mathbf{z}_i,$$

provided the partition associated to  $\mathbf{z}_1, \dots, \mathbf{z}_k$  is  $\pi$ ; otherwise, this probability equals zero.

- By definition of the extremal functions, one gets the joint law of the hitting scenario and max-stable observations:

$$\Pr\{\pi = \{\tau_1, \dots, \tau_k\}, Z(\mathbf{s}) = d\mathbf{z}\} = \exp\{-V(\mathbf{z})\} \left( \prod_{i=1}^k \int_{\mathbf{u}_i < \mathbf{z}_{\tau_i^c}} \lambda(\mathbf{z}_{\tau_i}, \mathbf{u}_i) d\mathbf{u}_i \right) d\mathbf{z}. \quad (11)$$

- By integrating out the hitting scenario, one obtains the law of the max-stable observations:

$$\Pr\{Z(\mathbf{s}) = d\mathbf{z}\} = \exp\{-V(\mathbf{z})\} \sum_{\pi \in \mathcal{P}_D} \left( \prod_{i=1}^k \int_{\mathbf{u}_i < \mathbf{z}_{\tau_i^c}} \lambda(\mathbf{z}_{\tau_i}, \mathbf{u}_i) d\mathbf{u}_i \right) d\mathbf{z}. \quad (12)$$

Equation (11) provides an alternative formula for the Stephenson–Tawn likelihood,  $g_{\text{ST}}$ , based on the Poisson point process intensity,  $\lambda$ , while Equation (12) is the max-stable full likelihood,  $g_{\text{Full}}$ . Identifying the expressions (11) and (12) above with (5) and (4) in the main paper, respectively, we can see that

$$-\partial_{\tau_i} V(z_1, \dots, z_D) = \int_{\mathbf{u}_i < \mathbf{z}_{\tau_i^c}} \lambda(\mathbf{z}_{\tau_i}, \mathbf{u}_i) d\mathbf{u}_i. \quad (13)$$

This relates a partial derivative of the exponent measure  $V$  with a partial integral of the point process intensity  $\lambda$ . In particular, (13) implies that the intensity is the mixed derivative of the exponent measure with respect to all arguments, i.e.,

$$\lambda(z_1, \dots, z_D) = -\frac{\partial^D}{\prod_{i=1}^D \partial z_i} V(z_1, \dots, z_D). \quad (14)$$

Furthermore, the function  $V$  corresponds to the integrated intensity of the set  $A = [0, \mathbf{z}]^c$ , i.e.,

$$V(z_1, \dots, z_D) = \Lambda([0, \mathbf{z}]^c) = \int_A \lambda(\mathbf{u}) d\mathbf{u}. \quad (15)$$

## A.2 Computing the Poisson point process intensity

The intensity measure  $\lambda$  is an important feature of max-stable models and can be computed for most popular models; see [Dombry \*et al.\* \(2013\)](#) for a derivation of  $\lambda$  for the Brown–Resnick model ([Kablichko \*et al.\*, 2009](#)) and [Ribatet \(2013\)](#) for an expression of  $\lambda$  for the extremal- $t$  model ([Opitz, 2013](#)). Partial integrals of  $\lambda$  for these models may be found in [Wadsworth and Tawn \(2014\)](#) and [Thibaud and Opitz \(2015\)](#), respectively. Using the relations (13) and (14), the intensity  $\lambda$  and its partial integrals can be deduced for the [Reich and Shaby \(2012\)](#) model from the expressions in the appendix of [Castruccio \*et al.\* \(2016\)](#).

Here, as a simple pedagogical illustration for many other multivariate or spatial max-stable models, we consider the multivariate logistic model, which we used in our simulation study. Notice that in this case, the function  $V$  and its partial and full derivatives can readily be obtained by direct differentiation.

Recall that the exponent measure for the logistic model is

$$V(z_1, \dots, z_D) = \left( z_1^{-1/\theta} + \dots + z_D^{-1/\theta} \right)^\theta, \quad \theta \in (0, 1].$$

It is known that the multivariate counterpart of the spectral representation (9) for the logistic model is obtained by taking  $W = (W_1, \dots, W_D)^T$  with independent and identically distributed Fréchet( $\beta, c_\beta$ ) components, where  $\beta = 1/\theta$  and  $c_\beta = 1/\Gamma(1 - 1/\beta)$  are shape and scale parameters, respectively; see, for example, Proposition 6 in [Dombry et al. \(2016\)](#).

Then,

$$f_W(z_1, \dots, z_D) = \prod_{i=1}^D \frac{\beta}{c_\beta} \left( \frac{z_i}{c_\beta} \right)^{-1-\beta} e^{-(z_i/c_\beta)^{-\beta}},$$

and we deduce from Equation (10) that

$$\begin{aligned} \lambda(z_1, \dots, z_D) &= \int_0^\infty \left[ \prod_{i=1}^D \frac{\beta}{c_\beta} \left( \frac{z_i}{rc_\beta} \right)^{-1-\beta} e^{-\{z_i/(rc_\beta)\}^{-\beta}} \right] r^{-2-D} dr \\ &= \frac{\Gamma(D-1/\beta)}{\beta} \left\{ \sum_{i=1}^D (z_i/c_\beta)^{-\beta} \right\}^{1/\beta-D} \prod_{i=1}^D \frac{\beta}{c_\beta} \left( \frac{z_i}{c_\beta} \right)^{-1-\beta}. \end{aligned}$$

Similar computations entail

$$\begin{aligned} \int_{\mathbf{u}_i < \mathbf{z}_{\tau_i^c}} \lambda(\mathbf{z}_{\tau_i}, \mathbf{u}_i) d\mathbf{u}_i &= \int_0^\infty \left[ \prod_{j \in \tau_i} \frac{\beta}{c_\beta} \left( \frac{z_j}{rc_\beta} \right)^{-1-\beta} e^{-\{z_j/(rc_\beta)\}^{-\beta}} \right] \times \left[ \prod_{j \in \tau_i^c} e^{-\{z_j/(rc_\beta)\}^{-\beta}} \right] r^{-2-|\tau_i|} dr \\ &= \left\{ \prod_{j \in \tau_i} \frac{\beta}{c_\beta} \left( \frac{z_j}{c_\beta} \right)^{-1-\beta} \right\} \int_0^\infty e^{-\sum_{j=1}^D \{z_j/(rc_\beta)\}^{-\beta}} r^{\beta|\tau_i|-2} dr \\ &= \frac{\Gamma(|\tau_i|-1/\beta)}{\beta} \left\{ \sum_{j=1}^D (z_j/c_\beta)^{-\beta} \right\}^{1/\beta-|\tau_i|} \prod_{j \in \tau_i} \frac{\beta}{c_\beta} \left( \frac{z_j}{c_\beta} \right)^{-1-\beta} \\ &= \beta^{|\tau_i|-1} \frac{\Gamma(|\tau_i|-1/\beta)}{\Gamma(1-1/\beta)} \left( \sum_{j=1}^D z_j^{-\beta} \right)^{1/\beta-|\tau_i|} \prod_{i \in \tau_i} z_j^{-1-\beta}, \end{aligned}$$

where, for the first equality, we used

$$\prod_{j \in \tau_i^c} \int_0^{z_j} \frac{\beta}{c_\beta} \left( \frac{u_j}{rc_\beta} \right)^{-1-\beta} e^{-\{u_j/(rc_\beta)\}^{-\beta}} du_j = \prod_{i \in \tau_i^c} r e^{-\{z_j/(rc_\beta)\}^{-\beta}}.$$



# References

- Booth, J. G. and Hobert, J. P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61**(1), 265–85.
- de Carvalho, M. and Davison, A. C. (2014) Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109**(506), 764–76.
- Castruccio, S., Huser, R. and Genton, M. G. (2016) High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Statist.* To appear.
- Davis, R. A., Klüppelberg, C. and Steinkohl, C. (2013) Statistical inference for max-stable processes in space and time. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**(5), 791–819.
- Davison, A. C. and Gholamrezaee, M. M. (2012) Geostatistics of extremes. *Proc. Roy. Soc. Edinburgh Sect. A* **468**(2138), 581–608.
- Davison, A. C. and Huser, R. (2015) Statistics of extremes. *Annu. Rev. Stat. Appl.* **2**, 203–35.
- Davison, A. C., Padoan, S. and Ribatet, M. (2012) Statistical modelling of spatial extremes (with discussion). *Statist. Sci.* **27**(2), 161–86.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39**(1), 1–38.
- Dombry, C., Engelke, S. and Oesting, M. (2016) Exact simulation of max-stable processes. *Biometrika* **103**(2), 303–17.
- Dombry, C. and Éyi-Minko, F. (2013) Regular conditional distributions of continuous max-infinitely divisible random fields. *Electron. J. Probab.* **18**(7), 1–21.
- Dombry, C., Éyi-Minko, F. and Ribatet, M. (2013) Conditional simulation of max-stable processes. *Biometrika* **100**(1), 111–24.

- Embrechts, P., Koch, E. and Robert, C. Y. (2015) Space-time max-stable models with spectral separability. *arXiv:1507.07750v1*.
- de Fondeville, R. and Davison, A. C. (2016) High-dimensional peaks-over-threshold inference for the Brown–Resnick process. *arXiv:1605.08558v1*.
- Genton, M. G., Ma, Y. and Sang, H. (2011) On the likelihood function of Gaussian max-stable processes. *Biometrika* **98**(2), 481–8.
- de Haan, L. (1984) A spectral representation for max-stable processes. *Ann. Probab.* **12**(4), 1194–204.
- de Haan, L. and Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. New York: Springer. ISBN 9780387239460.
- Huser, R. and Davison, A. C. (2013) Composite likelihood estimation for the Brown–Resnick process. *Biometrika* **100**(2), 511–8.
- Huser, R. and Davison, A. C. (2014) Space-time modelling of extreme events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**(2), 439–61.
- Huser, R., Davison, A. C. and Genton, M. G. (2016) Likelihood estimators for multivariate extremes. *Extremes* **19**(1), 79–103.
- Huser, R. and Genton, M. G. (2016) Non-stationary dependence structures for spatial extremes. *J. Agric. Biol. Environ. Stat.* To appear.
- Kabluchko, Z., Schlather, M. and de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *Ann. Probab.* **37**(5), 2042–65.
- Nielsen, S. F. (2000) The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6**(3), 457–89.

- Opitz, T. (2013) Extremal  $t$  processes: elliptical domain of attraction and a spectral representation. *J. Multivariate Anal.* **122**(1), 409–13.
- Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010) Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105**(489), 263–77.
- Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6**(4), 1430–51.
- Ribatet, M. (2013) Spatial extremes: Max-stable processes at work. *J. SFdS* **154**(2), 156–77.
- Schlather, M. (2002) Models for stationary max-stable random fields. *Extremes* **5**(1), 33–44.
- Shi, D. (1995) Fisher information for a multivariate extreme value distribution. *Biometrika* **82**(3), 644–9.
- Smith, R. L. (1990) Max-stable processes and spatial extremes. Unpublished.
- Stephenson, A. and Tawn, J. A. (2005) Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika* **92**(1), 213–27.
- Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C. and Heikkinen, J. (2016) Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. arXiv:1506.07836v1.
- Thibaud, E. and Opitz, T. (2015) Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102**(4), 855–70.
- Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statist. Sinica* **21**(2011), 5–42.
- Wadsworth, J. L. (2015) On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika* **102**(3), 705–11.

Wadsworth, J. L. and Tawn, J. A. (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101**(1), 1–15.